DOCUMENT RESUME

ED 250 331 TM 840 541

AUTHOR Kingston, Neal; Turner, Nancy
TITLE Analysis of Score Change Patterns of Examinees Repeating the Graduate Record Examinations General Test.
INSTITUTION Educational Testing Service, Princeton, N.J.; Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-84-22; GREB-83-5P
PUB DATE Jun 84
NOTE 49p.
AVAILABLE FROM Educational Testing Service, Research Publications R-116, Princeton, NJ 08541
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Entrance Examinations; Graduate Study; Higher Education; Multiple Regression Analysis; Quantitative Tests; Scores; Student Characteristics; *Test Format; *Test Reliability; Test Wiseness; Verbal Tests
IDENTIFIERS Analytical Tests; *Graduate Record Examinations; *Retesting

ABSTRACT

This investigation examines the impact the 1981 Graduate Record Examination (GRE) General Test Format Revision had on the stability over time of the verbal, quantitative, and analytical scores. Scores were used from the self-selected group of repeaters who took the GRE General Test twice between October 1980 and June 1982. Examinees were divided into three groups based on the test format taken. Three sets of analysis were performed: (1) analysis of difference scores and analysis of estimated true difference scores; (2) analysis of same form versus different form repeaters; and (3) analysis of predicted scores. Results for the new-format General Test show the greatest stability over time for the verbal measure and the least for the analytical measure. The large percentage of both gains and decreases on the analytical measure is primarily explained by the measure's lower reliability. Taking the same test edition twice provides the greatest expected score increase advantage for the analytical measure, and the least for the verbal measure. The psychometric characteristics with regard to score changes for the quantitative measure are between those for the verbal and analytical measures. Appendices include: (1) demographic characteristics of GRE repeaters; (2) group differences in background variables; and (3) regression equations. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED250331

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. C. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 840 511

GRE

GRADUATE RECORD EXAMINATIONS

ANALYSIS OF SCORE CHANGE PATTERNS OF EXAMINEES REPEATING THE GRADUATE RECORD EXAMINATIONS GENERAL TEST

Neal Kingston
Nancy Turner

GRE Board Professional Report GREB No. 83-5P
ETS Research Report 34-22

June 1984

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

Analysis of Score Change Patterns of Examinees Repeating the Graduate Record Examinations General Test[1]

Neal Kingston
Nancy Turner

GRE Board Professional Report GREB No. 83-5P

June 1984

[1]Major aspects of this research were designed by Nancy Burton and Madeline Wallmark. The consultation and review of numerous colleagues is gratefully acknowledged.

Copyright,© 1984 by Educational Testing Service. All rights reserved.

ABSTRACT

The stability of scores over time of the GRE® General Test verbal, quantitative, and analytical measures were studied using data from the self-selected group of repeaters. Overall, in these analyses the new-format GRE verbal measure demonstrated the greatest stability over time and the new-format analytical measure the least.

EXECUTIVE SUMMARY

In October 1981 the Graduate Record Examinations Program introduced a new version of the General Test that differed from the previous version in three major ways. A change in timing was designed to reduce the speededness of the verbal and analytical measures and to allow the addition of several quantitative items. Although the content of the verbal and quantitative measures remained essentially the same, the analytical measure was significantly revised by deleting two item types that had shown a short-term (within test form) practice effect and a susceptibility to coaching. Also, the analytical measure was rescaled. In addition to these changes, the scoring for all three measures was changed from "formula scoring" (subtracting a fraction of the number of incorrect responses from the number of questions answered correctly) to "rights scoring" (counting only the number of questions answered correctly). Instructions to the examinees were correspondingly changed: Examinees were advised to answer every question.

The purpose of this investigation was to examine the impact revising the format of the GRE® General Test had on the stability over time of the verbal, quantitative, and analytical scores. The investigation involved the use of scores from the self-selected group of repeaters who took the GRE General Test twice between October 1980 and June 1982. The intervals between administrations were either two, four, six, or eight months. These examinees can be divided into three groups: those who took the old-format test twice, those who took the new-format test twice, and those who took the old-format test first and the new-format test last. Some of the examinees who took the newformat test twice received the same edition of the test each time. Unless explicitly stated to the contrary, these same-edition repeaters were removed from all analyses.

The following major findings were noted.

- In general, the average score gain for repeaters taking the new-format GRE General Test twice is about 27 points on the verbal measure, 30 points on the quantitative measure, and 22 points on the analytical measure.
- Large gains, 100 or more points, are most common on the analytical measure (16% of all repeaters), next most common on the quantitative measure (13% of all repeaters), and least common on the verbal measure (9% of all repeaters).
- Large decreases, more than 100 points, are most common on the analytical measure (6% of all repeaters), next most common on the quantitative measure (2% of all repeaters), and least common on the verbal measure (1% of all repeaters).

- Even after the lower reliability of the analytical measure is taken into account, the variability of the difference scores for the analytical measure is slightly greater than for the quantitative measure and considerably greater than for the verbal measure. The standard deviations of the estimated true difference scores are 32 for verbal, 43 for quantitative, and 46 for analytical.

- On the new-format test, repeating the same edition of the test twice provided the greatest advantage on the analytical measure, next largest for quantitative, and smallest for verbal. For example, repeaters taking the same edition of the analytical measure two months apart averaged score gains of 47 points while those taking different editions had an average gain of only 17 points. The advantage for those taking the same edition was, therefore, 30 points. The advantages of taking the same edition of the verbal and quantitative measures two months apart were 13 and 20 points, respectively.

- On the new-format test, both different-edition repeaters and same edition repeaters averaged greater gains when they took the tests four months apart than when they took the tests two months apart.

- The average gain on the verbal measure is about five points more on the new-format test than it was on the earlier verbal measure. This difference is statistically significant beyond the .001 level.

- The average gain on the quantitative measure is about three points more on the new-format test than it was on the earlier quantitative measure. This difference is not statistically significant at any commonly accepted level.

- The average gain on the new-format analytical measure is about 28 points less than it was on the earlier analytical measure. This difference is statistically significant beyond the .0001 level. This is reassuring as the measure was revised to reduce short-term practice effect.

- The test-retest correlations, for both the earlier General Test and the new-format test, decrease as the time between administrations increases from two months to eight months.

- The test-retest correlations for the verbal measure are higher than those for the quantitative measure for both the old- and new-format General Tests.

- The test-retest correlations for the new-format analytical measure are considerably lower than any of the other test-retest correlations. This is partially accounted for by the lower internal consistency reliability of the analytical measure.

- For the new-format test, when the distributions of change scores are corrected for unreliability, using test-retest correlations to account for differences in reliability among the three measures, the analytical measure change scores were most variable. The variability of the quantitative measure change scores was only slightly lower, both being considerably higher than the variability of the verbal measure change scores.

In general, the analyses performed on the scores of the self-selected population of repeaters of the new-format General Test show that the verbal measure demonstrates the greatest stability over time and the analytical measure the least. The analytical measure has the greatest percentage of large gains and large decreases. These changes are primarily explained by the lower reliability of the measure. Taking the same edition twice provides the greatest advantage with regard to expected score increase for the analytical measure. The test-retest correlations for the analytical measure are considerably lower than those for the other measures. On the positive side, the average gain for repeaters taking different editions of the test is somewhat lower than that for the other two measures.

The verbal measure, on the other hand, has the smallest percentage of large score changes. Taking the same edition twice provides examinees with less advantage on the verbal measure than on the other two measures. Also, the test-retest correlations are the highest of the three measures. The psychometric characteristics with regard to score changes for the quantitative measure are between those for the verbal and analytical measures.

It should be noted that during the period in which this study was conducted, test forms had to be repeated more frequently than is the normal GRE policy. This was necessitated by the development and initial use of the new-format measures. The frequency of repetition has since decreased.

Table of Contents

Abstract 1

Executive Summary. ii

Introduction 1

Research Design. 3

Data Base. 3

Methods. 5

Results. 7

Conclusions. 21

References 23

Appendix 1 - Demographic Characteristics of GRE Repeaters. 25

Appendix 2 - Differences with Regard to Background Variables of Three Groups of Repeaters. 33

Appendix 3 - Regression Equations. 39

Tables and Figures

Table 1 - Definition of Study Sample 4

Table 2 - Distribution of Score Changes for Repeaters Who Took the New-Format GRE General Test Twice in 1981-82 7

Table 3 - Mean Score Change for Repeaters of the New-Format GRE General Test Taking the Same Edition or Two Different Editions 9

Table 4 - Means and Standard Deviations of the Difference Scores for the Old-Old and New-New Groups 10

Table 5 - Analysis of Covariance on Verbal Difference Scores 11

Table 6 - Analysis of Covariance on Quantitative Difference Scores . . 12

Table 7 - Analysis of Covariance on Analytical Difference Scores . . . 12

Table 8 - Means and Standard Deviations of the Estimated True Difference Scores for the Old-Old and New-New Groups 13

Table 9 - Test-Retest Correlations for the GRE General Test Repeaters (New and Old Formats) 14

Table 10 - Predicting Test Scores Across a Change in Test Format: Means and Standard Deviations of Residuals and Correlations of Predicted Scores with Observed Scores in Two Cross-Validation Groups 15

Table A1-1 - Frequencies and Adjusted Percentages of Repeaters and Nonrepeaters on Demographic Variables 28

Table A1-2 - Number of Repeaters Scoring 200 on Each Measure of the GRE General Test 30

Table A2-1 - Comparisons of Repeater Groups (New-New, Old-Old, Old-New) by Background Variables 36

Table A3-1 - Prediction Equation for Second Verbal Score 39

Table A3-2 - Prediction Equation for Second Quantitative Score 40

Figure 1 - Residual Bands of GRE Verbal Scores in the Old-New Group 19

Figure 2 - Residual Bands of GRE Quantitative Scores in the Old-New Group 20

INTRODUCTION

In October 1981 the Graduate Record Examinations Program introduced a new version of the General Test that differed from the previous version in three major ways. A change in timing was designed to reduce the speededness of the verbal and analytical measures and to allow the addition of several quantitative items. Although the content of the verbal and quantitative measures remained essentially the same, the analytical measure was significantly revised by deleting two item types that had shown a short-term (within test form) practice effect and a susceptibility to coaching. Also, the analytical measure was rescaled. In addition to these changes, the scoring for all three measures was changed from "formula scoring" (subtracting a fraction of the number of incorrect responses from the number of questions answered correctly) to "rights scoring" (counting only the number of questions answered correctly). Instructions to the examinees were correspondingly changed: Examinees were advised to answer every question.

As part of the effort to evaluate the impact of the restructuring on the psychometric properties of the General Test, test analysis reports for the five test forms introduced in 1981-82 were reviewed (Wallmark, 1982a-e). In addition, special analyses of the concurrent validity of the new-format test were conducted. A report summarizing these efforts to evaluate the new General Test (Goodison & Wild, 1982) was presented to the Graduate Record Examinations Board in September 1982. That report recommended a study of the large score gains and losses on the analytical measure, and this paper is a response to the GRE Board's request for such a study.

The purpose of this investigation, which is part of the larger effort to evaluate the new-format General Test, was to examine the stability over time of the verbal, quantitative, and analytical scores or the new-format General Test. This report will try to answer the following questions:

- What is the score change pattern for repeaters who took the new-format General Test in 1981-82?
- Do the score change patterns for the verbal, quantitative, and analytical measures differ?
- How do score changes for repeaters who took the same edition twice compare with the changes for those who took two different editions?
- Do the patterns of score changes on the new-format test (1981-82) differ from the patterns of score changes on the previous version of the test (1980-81)?
- Do any of the score change patterns indicate a change in the constructs underlying the verbal and quantitative measures?

An earlier study of repeaters (Rock & Werts, 1980) showed that the mean gain on the verbal measure tended to increase as the time between tests increased -- an indication that the test indeed measured skills that were developed over the long term. The average gain for the quantitative measure was, in general, smaller, and the effect of time between tests was less clear.

One important question about the new-format test is whether the change to rights scoring -- which may have encouraged guessing -- affected the previously established pattern of score changes.

The analytical measure is especially of interest since the major changes in the scale resulted in little available knowledge regarding the stability of scores over time. Preliminary data suggested that this stability was likely to be low relative to the verbal and quantitative measures. Because the item types that could be answered more quickly were dropped, the new-format test contains only 50 items. Also, the analytical measure is slightly more difficult than the verbal and quantitative measures. One disturbing consequence of its greater difficulty is that the analytical measure yields a larger proportion of very low scores than the other two measures -- scores that could have been obtained by random guessing alone. The resulting measure has had an estimated reliability below .90 for eight of the first ten forms.

RESEARCH DESIGN

Data base

The analyses are based on the self-selected group of Graduate Record Examinations test takers who took the General Test the first time between October 1980 and April 1982 and took the test at least one additional time by June 1982. Only those examinees taking the test in the pairs of domestic administrations indicated in Table 1 were included. Only the scores, verbal, quantitative, and analytical, and background information from their first two administrations were considered. It is likely that some repeaters were not identified, particularly those who changed their names or made substantial errors in gridding identification information on their answer sheets.

The sample can be broken down into three groups:

- (New-New) the GRE examinees who took the new-format General Test at least twice in the first operational year 1981-82 and have no previous scores on record;
- (Old-Old) the GRE examinees who took the previous version of the General Test at least twice in 1980-81 and have no General Test scores on record earned prior to October 1980; and
- (Old-New) the GRE examinees who took the General Test for the first time in 1980-81 and the second time in 1981-82 (i.e., the old format first and then the new format).

Table 1 indicates which pairs of administrations constitute each of the three samples, which forms were administered, and how many examinees took the pair of forms. The number of months between the administrations are indicated in parentheses. Pairs of administration for which identical forms were administered are indicated by an asterisk. It is a general practice, within the GRE program, not to administer the same form too frequently. The development and initial use of the new-format measures necessitated the extreme repetition of forms in the 1981-82 testing year. The frequency of repetition has since decreased.

Table 1

Definition of Study Sample[1]

New-New (n = 4,084)

First Administration		Date and Form of Second Administration					
Date	Form	12/81 3DGR1	2/82 3DGR2	4/82 K-3DGR3	4/82 3EGR1	6/82 3EGR1	6/82 3EGR2
10/81	3DGR1	253*(2)	277 (4)	27 (6)	102 (6)	40 (8)	28 (8)
10/81	3DGR2	231 (2)	251*(4)	25 (6)	79 (6)	28 (8)	31 (8)
10/81	3DGR3	267 (2)	262 (4)	19*(6)	117 (6)	38 (8)	29 (8)
12/81	3DGR1		494 (2)	92 (4)	329 (4)	102 (6)	93 (6)
2/82	3DGR2			75 (2)	325 (2)	132 (4)	111 (4)
4/82	K-3DGR3					26 (2)	24 (2)
4/82	3EGR1					86*(2)	91 (2)

Old-Old (n = 5,072)

First Administration		Date and Form of Second Administration			
Date	Form	12/80 K2-3AGR2	2/81 K-3BGR1	4/81 K-3BGR2	6/81 3CGR1
10/80	K-3AGR1	1,187 (2)	1,048 (4)	413 (6)	233 (8)
12/80	K2-3AGR2		495 (2)	485 (4)	230 (6)
2/81	K-3BGR1			445 (2)	239 (4)
4/81	K-3BGR2				297 (2)

Old-New (n = 2,453)

First Administration		Date and Form of Second Administration				
Date	Form	10/81 3DGR1	10/81 3DGR2	10/81 3DGR3	12/81 3DGR1	2/82 3DGR2
2/81	K-3BGR1	183 (8)	100 (8)	101 (8)		
4/81	K-3BGR2	201 (6)	227 (6)	231 (6)	307 (8)	
6/81	3CGR1	239 (4)	246 (4)	238 (4)	261 (6)	119 (8)

[1]Tabled entries indicate number of examinees who were administered the pair of editions and, in parentheses, the number of months between administrations. An asterisk indicates that examinees took the same edition both times.

Methods

Data were analyzed by several different methods. Distributions and summary statistics of difference scores were calculated for verbal, quantitative, and analytical scores for the New-New and Old-Old groups, not including examinees who took the same edition twice. This analysis is referred to as analysis of difference scores. Distributions and summary statistics were also calculated for estimated true difference scores using a method developed by Lord (Lord & Novick, 1968). This method is referred to as analysis of estimated true difference scores. To some extent the assumptions of the method were violated[1] in this analysis and the effects of these violations are not clear. The statistical significance of difference between the New-New and Old-Old distributions for each section was assessed using an analysis of covariance.

A second set of analyses was based on a subset of the New-New group: those examinees who were first tested in October 1981 and tested second in either December 1981 or February 1982. This is referred to as analysis of same form vs different form repeaters. Three different editions of the GRE General Test were administered in October 1981, and equivalent groups of examinees[2] took each edition. One of these editions was repeated in December 1981, and thus an essentially random one-third of the examinees in the 10/81 - 12/81 sample took the same edition twice and the remaining two-thirds (those who took one of the other two editions in October) took different editions. Likewise, another edition that was administered in October 1981 was repeated in February 1982, providing another pair of samples in which one-third took the same edition and two-thirds took different editions. For each scale, the mean difference scores, standardized mean difference scores[3], and standard deviations of the mean difference scores for those examinees taking the same edition and those taking different editions were compared. The examinees who took the same edition of the test both times were not included in any of the other analyses.

The third set of analyses made use of the data collected from examinees in the Old-New group. For each measure, using data from the Old-Old group, stepwise regression was used to predict second scores from first scores, functions of first scores (e.g., first score squared), and demographic data. This method is referred to as analysis of predicted scores. Appendix 3 lists the variables used and their weights.

---

[1]Most notably, the method assumed a linearity of the regression of true difference on observed first and second scores when it is certain that this regression is, at least to some extent, nonlinear.

[2]The three test editions were packaged in alternating order and assigned to examinees in seating order. Past experience indicates samples formed this way are essentially random within test center and stratified by test center.

[3]The number of standard deviations of the difference score that the mean difference score is from zero.

The resulting prediction equations were cross-validated twice. Equations were cross-validated in a random holdout sample from the Old-Old group and in the Old-New group. Assuming that the Old-New cross-validation group was a random sample, any differences between the distributions of predicted scores or the distributions of residuals would be a reflection of a change in scale, a change in the construct(s) underlying the scale, or sampling error. Even if this assumption is violated, to the extent that all variables that accounted for significant differences between the samples were in the regression analyses, differences in the distributions of predicted score residuals can be interpreted as if the assumption held.

RESULTS

At the beginning of this paper, five questions were raised regarding score changes for GRE repeaters. In this section, those questions will be raised again and answered explicitly.

- What is the score change pattern for repeaters who took the new-format General Test in 1981-82?

Table 2 presents the score change pattern for repeaters who took the new-format General Test. This table consists of the frequency distributions of the change scores, grouped into 50-point intervals, for the verbal, quantitative, and analytical measures. At the bottom of the table are the mean and standard deviation of the score changes for each of the three measures. In addition, the percent of repeaters whose scores increased considerably (100 or more points), decreased considerably (more than 100 points), or did not change very much (-100 to + 99 points) is indicated for each group of scores.

Table 2

Distribution of Score Changes for Repeaters
Who Took the New-Format GRE General Test Twice in 1981-82
(N = 3,475)

Score Change	Verbal		Quantitative		Analytical	
250+	3		11		13	
200 to 240	13	9%	17	13%	27	16%
150 to 190	46		93		125	
100 to 140	242		334		386	
50 to 90	735		741		669	
0 to 40	1201	90%	1052	84%	807	78%
-50 to -10	984		857		838	
-100 to -60	214		284		405	
-150 to -110	32		72		144	
-200 to -160	5	1%	12	2%	46	6%
-250 to -210	0		1		11	
-260-	0		1		4	
Mean	26.6		29.5		22.0	
Standard Deviation	58.4		69.1		82.5	

- Do the score change patterns for the verbal, quantitative, and analytical measures differ?

The mean score changes, as indicated in Table 2, are 26.6 for verbal, 29.5 for quantitative, and 22.0 for analytical. While, given the large sample size, these differences would surely be statistically significant, it should be remembered that neither equal scores nor equal score changes on the three General Test measures indicate the same relative performance.[1]

Although the observed means differ somewhat, the most striking feature of Table 2 is the difference in the variability of the three measures. The standard deviations of the change scores, as indicated in Table 2, are 58.4 for verbal, 69.1 for quantitative, and 82.5 for analytical. The differences among the three measures are even more striking with regard to the percent of repeaters whose scores increase or decrease considerably. On the verbal measure, only 9 percent of the repeaters had score increases of 100 points or more, and only 1 percent had decreases of more than 100 points. On the quantitative measure, 13 percent of the repeaters had increases of 100 points or more and 2 percent had decreases of more than 100 points. Sixteen percent of the repeaters had score gains of 100 points or more on the analytical measure, almost twice as many as for the verbal measure. Six percent of the repeaters had score decreases of more than 100 points, about six times as many as for the verbal measure. The direction of these findings is in keeping with the differences in the reliabilities of the three measures, with the verbal measure being most reliable and the analytical measure being least reliable.

- How do score changes for repeaters who took the same edition twice compare with those who took two different editions?

Table 3 presents the mean score changes on the new-format General Test for two kinds of repeaters: those who took the same form of the test twice and those who took two different forms of the test. This sample consists of 1,541 examinees chosen so that, for both the two-month and four-month groups, the group who took the same form and the group who took different forms are essentially random subsamples.

---

[1]The verbal and quantitative measures were scaled in 1952 to have the same mean and standard deviation for a selected group of college seniors. The analytical measure was scaled so that its mean and standard deviation would be between the corresponding statistics for verbal and quantitative scores for the group of examinees tested in October 1981. As the makeup of the GRE population changed from that of the scaling groups, the scores of the three measures became less comparable. Also, the change scores were never scaled to be comparable.

Table 3

Mean Score Change for Repeaters of the New-Format GRE General Test Taking the Same Edition or Two Different Editions (N = 1,541)

	Sample Size	Mean Score Change V	Q	A	Score Change S.D. V	Q	A
2 Months Between Administrations							
Same Form	253	36	36	47	50	58	70
Different Forms	498	23	16	17	63	65	78
Difference		13	20	30			
4 Months Between Administrations							
Same Form	251	40	46	49	55	60	69
Different Forms	539	31	31	29	58	74	79
Difference		9	15	20			

Table 3 shows that for repeaters with two months between administrations, the advantage of taking the same edition twice was greatest for the analytical measure and least for the verbal measure. For repeaters with four months between administrations, the results were similar. Note that the examinees who repeated the same edition of the analytical measure after two months gained nearly as much as those who repeated it after four months. On the quantitative measure, the four-month same-edition repeaters gained 10 points more than the two-month repeaters. For examinees taking different editions of the test, the differences between two and four months are largest for quantitative and smallest for verbal.

- Do the patterns of score changes on the new-format test (1981-82) differ from the patterns of score changes on the previous version of the test (1980-81)?

Table 4 presents three types of mean difference scores, the standard deviations of the differences, and the sample sizes, both for repeaters who took the old-format GRE General Test twice and those who took the new-format test twice. The three mean difference scores are observed, adjusted with an analysis of covariance, and standardized -- the observed difference divided by the standard deviation of the differences. The standard deviations of the differences are given in the last column. Tables 5, 6, and 7 present the analyses of covariance for the verbal, quantitative, and analytical measures.

Table 4

Means and Standard Deviations of the Difference Scores for the Old-Old and New-New Groups[1]

Scale	Group	Sample Size	Type of Mean Score: Observed	Adjusted[2]	Standardized[3]	Standard Deviation
V	Old-Old	5072	22.2	22.1	.40	56.2
	New-New	3475	26.6	26.7	.46	58.4
Q	Old-Old	5072	26.2	26.5	.40	65.5
	New-New	3475	29.5	29.1	.43	69.1
A	Old-Old	5072	41.0	40.6	.59	70.0
	New-New	3475	22.0	22.4	.27	82.5

The standard deviations of the differences show that, for the old-format measures, the verbal difference scores were least variable and the analytical difference scores were most variable. After the change in format, the verbal difference scores remained about as variable as they had been, the quantitative difference scores were slightly more variable, and the analytical difference scores became considerably more variable. The increase in the variability of the analytical difference scores is in keeping with the lower reliability of the new-format analytical measure.

Table 5 shows that, for the verbal measure, the disparity between the old-format mean difference and the new-format mean difference is statistically significant. Table 4 shows that this statistically significant difference between the difference scores is fairly small; that is, on the new format verbal measure the average observed gain for repeaters was 4.4 points more than the average gain on the old-format verbal measure, and the average covariance-adjusted gain was 4.6 points more. The next to the last column shows that the difference between the standardized difference scores is only .06 standard deviations.

---

[1] Old-Old indicates that the examinees took the old-format GRE General Test both the first and the second time they took the test. New-New indicates that they took the new-format test both the first and the second time.

[2] Scores were adjusted with an analysis of covariance. The covariates are indicated in the analysis of covariance tables.

[3] Number of standard deviations of the difference score from zero.

Table 4 shows that the difference between gains for the old- and new-format quantitative test is small: The difference between average observed gains was 3.3 points, and the difference between average covariance-adjusted gains was 2.6 points. The gain, in standard deviations of the difference scores, was .03. The analysis of covariance in Table 6 shows that these differences are not statistically significant at any commonly accepted level.

For the analytical score, Table 4 shows that the difference between average observed gains was 19.0 points, and the difference between average covariance-adjusted gain was 18.2 points. The observed difference translated to .32 standard deviations of the difference score. Table 7 shows that these differences are statistically significant at beyond the .0001 level. This is reassuring, as the large repeater gain on the old-format analytical measure is primarily due to practice effect on analysis of explanations and logical diagrams items (Kingston & Dorans, 1982; Swinton, Wild, & Wallmark 1983). Obviously, the attempt to reduce this practice effect, which was instituted after these studies, was successful.

Table 5

Analysis of Covariance on Verbal Difference Scores

Source of Variance	Sum of Squares	df	Mean Square	F Ratio	p
Between Formats	40,013	1	40,013	12.31	.0005
Covariates	147,638	4	36,910	11.36	.0000
Error	27,760,536	8,541	3,250		

Covariates and Their Weights[1]

Ethnic: White	5.9
Degree objective: Master's	-4.6
Graduate major: Education	-7.7
Months between administrations[2]	1.0

[1]Covariates are dummy coded 0 if the examinee was not in the group and 1 if the examinee was in the group.

[2]Months between the two administrations: either 2, 4, 6, or 8.

Table 6

Analysis of Covariance on Quantitative Difference Scores

Source of Variance	Sum of Squares	df	Mean Square	F Ratio	p
Between Formats	10,861	1	10,861	2.44	.1182
Covariates	318,963	4	79,741	17.93	.0000
Error	37,979,543	8,541	4,447		

Covariates and Their Weights[1]

English primary language	-7.5
Sex: Female	-5.3
UGPA: B-	-5.8
Months between administrations[2]	2.2

Table 7

Analysis of Covariance on Analytical Difference Scores

Source of Variance	Sum of Squares	df	Mean Square	F Ratio	p
Between Formats	635,542	1	635,542	112.15	.0000
Covariates	171,341	3	57,114	10.08	.0000
Error	48,405,363	8,542	5,667		

Covariates and Their Weights[1]

U.S. citizen	8.5
Resident alien	.6
Ethnic: Black	-11.7

[1]Covariates are dummy coded 0 if the examinee was not in the group and 1 if the examinee was in the group.

[2]Months between the two administrations: either 2, 4, 6, or 8.

In summary, Tables 4 through 7 show that, on the average, for the new-format General Test, repeaters gain a little more on the verbal measure but much less on the analytical measure than they did on the old-format General Test.

Table 8

Means and Standard Deviations of the Estimated True Difference Scores for the Old-Old and New-New Groups

Scale	Group	Sample Size	Mean	Standard Deviation
V	Old-Old	5072	22.2	32.2
	New-New	3475	26.6	31.9
Q	Old-Old	5072	26.2	29.9
	New-New	3475	29.5	43.1
A	Old-Old	5072	41.0	37.5
	New-New	3475	22.0	45.5

Table 8 presents the means and standard deviations of the estimated true difference scores in a format similar to Table 4. The estimated mean true difference scores are the same as the observed mean difference scores, as the observed difference scores were all shrunk to the observed mean, and this shrinking was symmetrical. The standard deviations, however, were greatly influenced by this procedure. Not only are the standard deviations of the estimated true difference scores much smaller than the observed standard deviations, but their relative magnitudes have been changed. Once the differing test-retest reliabilities of the measures have been taken into account by this analysis, it can be seen that the remaining variability of the difference scores is larger for the new-format quantitative and analytical measures than it was for the respective old-format measures. The large difference between standard deviations for the quantitative measure is surprising, as the changes made to that measure appear relatively minor.

Table 9

Test-Retest Correlations
for GRE General Test Repeaters
(New and Old Formats)

Scale	New or Old	Same or Different	Median Reliability[1]	Correlations[2] Months Between Admin. 2	4	6	8
V	Old	Different	.92	.90	.88	.88	.83
	New	Different	.92	.88	.86	.86	---
	New	Same	.92	.93	.89	---	---
Q	Old	Different	.90	.86	.84	.84	.83
	New	Different	.91	.84	.81	.83	---
	New	Same	.91	.91	.86	---	---
A	Old	Different	.93	.86	.83	.84	.81
	New	Different	.86	.76	.73	.69	---
	New	Same	.86	.85	.82	---	---
Sample Size	Old	Different	2424	1772	643	233	
	New	Different	1533	1203	545	194	
	New	Same	339	251	19	0	

Table 9 presents the test-retest correlations for examinees taking two different editions of the old-format GRE General Test, for examinees taking two different editions of the new-format test, and for examinees taking the same edition of the new-format test twice. For comparison purposes reloability estimates based on test analyses samples (sample sizes usually about 2,000) are included. These data are probably, to some extent, overestimates for the population of repeaters, as repeater scores are less variable than the general GRE test-taker population scores. Reliability was estimated as:

$$1 - \frac{\text{error variance}}{\text{total variance}},$$

where error variances were estimated, using KR-20, separately for each item type within each section and summed across all item types in the measure (Wallmark 1982a-e).

---

[1]Reliability calculated as described in text.

[2]Correlations based on samples of less than 200 are not reported.

The first feature of note in Table 9 is that the median reliability for the new-format analytical measure is much lower than that for the old-format analytical measure, and, in fact, is much lower than that for the old and new formats of the verbal and quantitative measures. This finding is well known and has been presented previously elsewhere (Kingston, 1984; Wallmark, 1982a-e). A second noteworthy feature is that, not surprisingly, the test-retest correlations for all measures become lower as the time between administrations increases. Third, correlations were consistently higher when examinees took the same edition of the test twice rather than two different editions. And, last, the correlations for the new-format measures when repeaters took two different editions of the test were consistently slightly lower than for the old-format measures when repeaters took two different editions.

Table 10 presents the means and standard deviations of the residuals and the correlations of predicted and observed second test administration scores for those examinees in two cross-validation groups: the first, a random holdout sample from the Old-Old group from which the prediction equations were derived, the second, the Old-New group. Old-Old cross-validation samples for the analyses predicting second scores serve as baselines for comparing the residuals in the Old-New cross-validation samples. Although we know that the expected residual in a randomly selected cross-validation group will be zero, these data allow us to get a handle on how much deviation from the expected zero is reasonable and also provide us with an estimate of the expected standard deviation of the residuals and the expected multiple correlation. As the changes in the analytical measure were not intended to maintain the underlying constructs of the measure nor maintain the score scale, these analyses are presented only for the verbal and quantitative measures.

Table 10

Predicting Test Scores Across a Change in Test Format: Means and Standard Deviations of Residuals and Correlation of Predicted Scores with Observed Scores[1] in Two Cross-Validation Groups

		V			Q		
Format	Sample Size	$\bar{x}$	s	r	$\bar{x}$	s	r
Old-Old	997	1.2	53.7	.90	-.8	62.4	.86
Old-New	2353	-5.1	54.1	.89	-9.4	63.2	.87
New-New	1022	3.5	53.8	.88	1.3	62.4	.85
Old-New	2353	12.2	52.5	.89	10.2	60.4	.87

[1]For comparison, the multiple correlations in the Old-Old sample (n = 4,075) in which the prediction equations were calculated were .90, .87, and .88 for verbal, quantitative, and analytical, respectively.

In predicting the second administration scores, the verbal mean residual in the Old-New group appears slightly low, and the quantitative mean residual appears moderately low. The standard deviations of the residuals and the correlations of predicted scores with observed scores for the verbal and quantitative measures appear in line with those from the holdout cross-validation group.

- Do any of the score change patterns indicate a change in the constructs underlying the verbal and quantitative measures?

Scrutiny of the residuals from the predicted score analyses provides evidence regarding possible change in the constructs underlying the scales. Figures 1 and 2 present plots of the mean residuals for each 50-point interval of predicted score: 200-250, 250-300, and so on. No data are presented for any score grouping for which there are fewer than 30 predicted scores. These points are joined by a cubic spline fit to provide an estimate of the mean residuals throughout the predicted score range. Two additional lines surround the line of mean residuals. The upper line is the cubic spline fit connecting the points that are one standard deviation (based on the data within each individual score grouping) above the mean residuals. The lower line is like the upper, only connecting the points one standard deviation below the mean residuals.

If scores on the old-format test and the new-format test were completely interchangeable, one would expect that for each measure the mean residuals would be zero throughout the predicted score range, and the standard deviation of the residuals would be constant. Minor perturbations might be due to sampling error. More serious deviations might be caused by one or more of three factors:

1. one or more important variables were left out of the regression analyses;

2. the original equating of the new-format test to the old-format test was not totally appropriate; or

3. the construct(s) underlying the measures have changed.

It is unlikely that important variables were left out of the regression analyses. Data on more than 100 potentially useful variables were input into the regression. There was little increment in variance explained beyond the fourth or fifth predictor selected, yet 13, 22, and 20 predictors were selected as statistically significant for the verbal, quantitative, and analytical measures respectively. Since GRE annual populations are fairly stable from year to year, and the percentile ranks for comparable verbal and quantitative scores in 1980-81 and 1981-82 matched very closely, it is very unlikely that there could be more than one or two scaled score points of error in the equating of those two measures across the change in test format. Thus, it is likely that most of the variation from the expected pattern of residuals

is due to changes in the underlying constructs. Since the content of the verbal and quantitative measures did not change, the change in construct is likely to be related to the change in speededness or the change to rights-scoring instructions.

Figure 1 shows that for the verbal measure, the mean residuals were positive for the lower half of the predicted score range (200-400) and negative for the upper half (450-700). Only at the lowest part of the score range (200-250) was the mean residual more than slightly positive. The figure shows, for example, that examinees for whom we predicted scores between 600 and 700, on the average, scored 20 to 25 points lower than predicted. throughout most of the predicted score range, the mean residual was strongly linearly related to the predicted score. The standard deviations of the residuals are very consistent (about 52-56) throughout most of the score scale (300-700). At the lower end of the scale (200-300) they are slightly lower (48-49). All in all, the residual pattern for the predicted verbal scores indicates a small but consistent change in the sources of variability underlying the verbal measure.

The three major changes in the verbal measure were the shortening of the test from 80 to 76 items, increasing the time limit from 50 to 60 minutes, and changing from formula-scoring instructions to right-scoring instructions. The first two changes were instituted to decrease the speededness of the test. This might result in a pattern of residuals like the one in Figure 1. Test items are, in general, ordered by difficulty, and so typically only the more able examinees (those who might have been able to answer some of the last few difficult items) earn scores that are affected by speededness. On the less speeded new-format verbal measure, the more able students would be more likely to have had their scores underpredicted, and thus they would be more likely to have negative residuals.

Alternatively (or in addition), a pattern like the one observed might be the result of a greater propensity toward guessing under rights-scoring instructions for higher scoring examinees. This propensity might be due to lower scoring examinees, particularly nonnative English speakers, being less able to read and understand the instructions. Alternatively, some lower scoring examinees might be less able to break any existing psychological set against guessing that might have been imposed by previous GRE test-taking experiences under formula-scoring instructions.

Figure 2 presents the residual bands for the quantitative measure, based on the prediction equation for second scores derived in the Old-Old group. Throughout the middle of the predicted score range (400-650) the mean residuals are slightly negative (about 5 to 10 points). At both ends of the scale (250 to 400 and 650 to 750) the mean residuals are moderately negative (10 to 20 points). The standard deviations of the residuals were largest (69-70) between predicted scores of 350 and 450, and decreased toward the high end of the scale (62 from 500 to 600, 56 from 650 to 700, and 49 from 700 to 750).

Overall, the residual pattern for the quantitative measure shows consistent overprediction of quantitative scores on the new-format General Test. The only substantial change in the quantitative measure was the change to right-scoring instructions. If the self-selected group of repeaters was less likely to guess under the old format formula-scoring instructions than typical examinees, but under rights-scoring directions was just as likely to guess, a residual pattern like the one observed would have occurred. Assuming that many repeaters undergo additional studying or coaching before taking the test a second time, it is possible that they would learn to supress any reticence to guess they had, particularly under circumstances where they could eliminate one or more of the distractors. This hypothesis does not explain why the standard deviation of the residuals would be lower for the higher predicted scores.

Figure 1

Residual Bands of GRE Verbal Scores in the Old-New Group

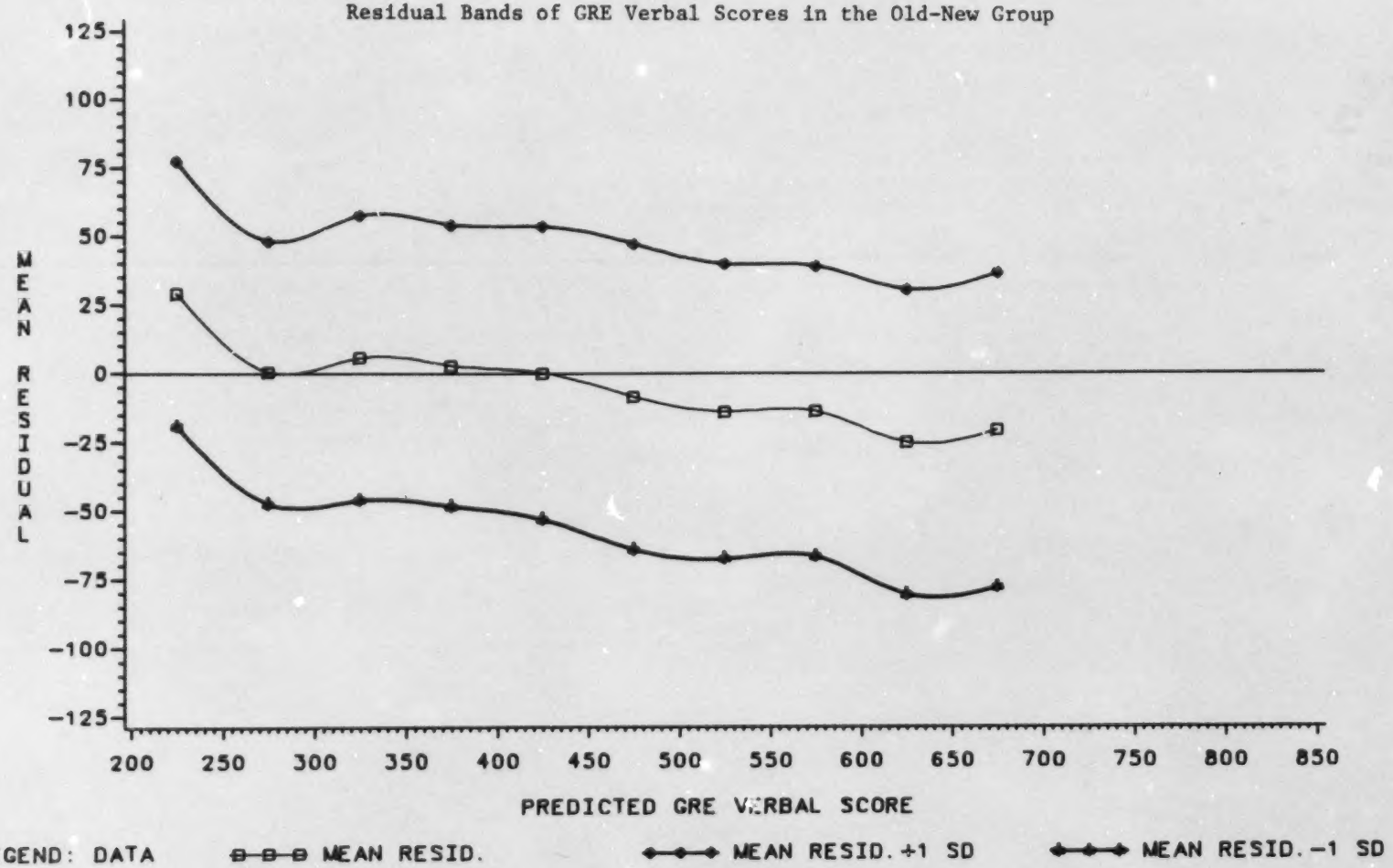
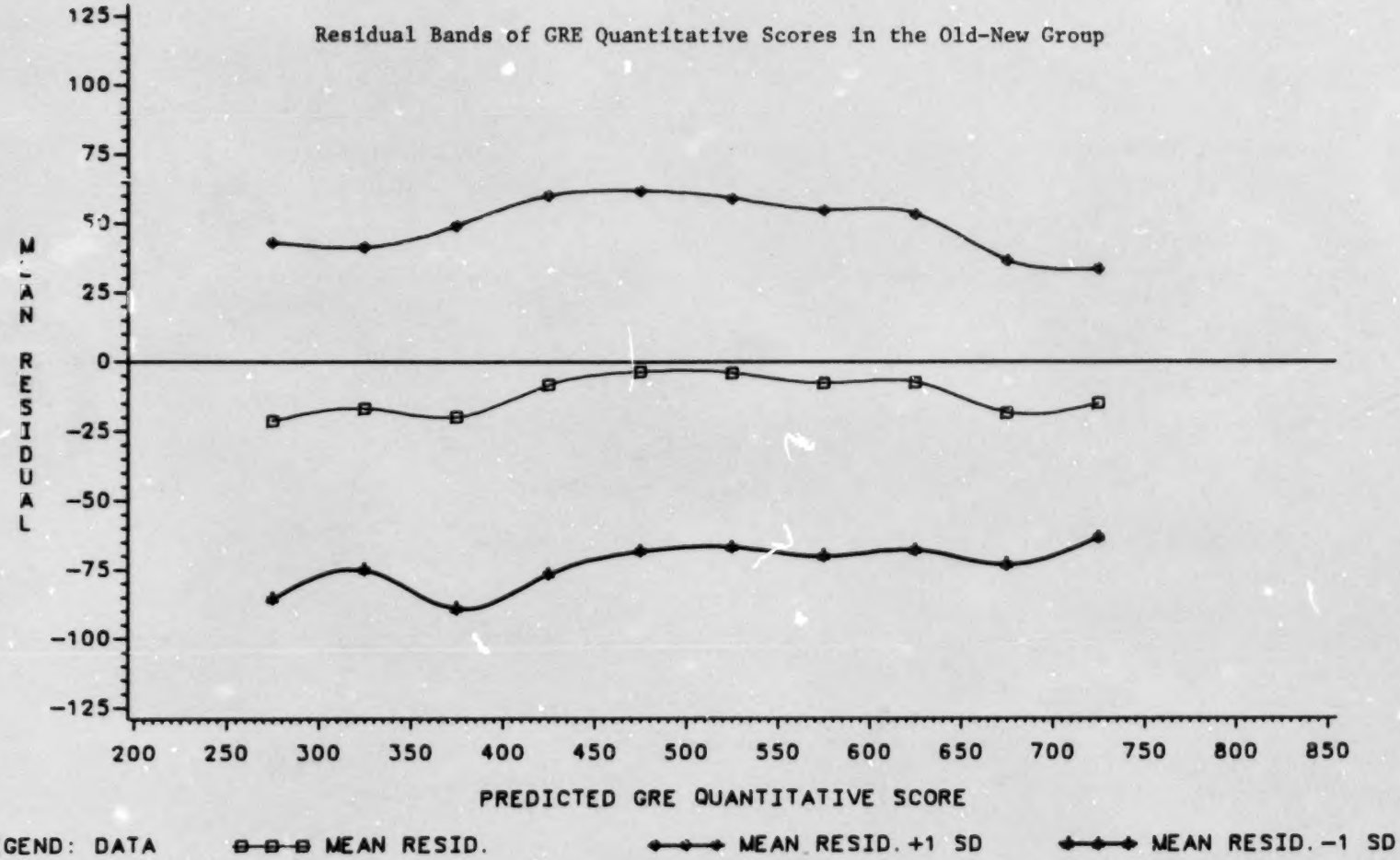


Figure 2

Residual Bands of GRE Quantitative Scores in the Old-New Group



CONCLUSIONS

In interpreting the results of this study, it must be remembered that the data are based on the self-selected population of repeaters, and as such, the results would be likely to differ if the study were performed on a randomly selected group of General Test examinees. Also, since the analytical measure is known to be experimental, few examinees might choose to repeat the General Test because of dissatisfaction with their analytical score.

The average score gain for repeaters is not very different for the three General Test measures, ranging from 22 points to 30 points. Also, for the sample of examinees who took the same edition of the test twice, the average gains were similar for the three measures, ranging from 36 points to 47 points with two months between administrations and 40 points to 49 points for four months between administrations. In terms of other characteristics, the three measures varied considerably.

In general, the analyses performed show that with regard to stability of scores over time for the self-selected population of repeaters taking the new-format General Test, the verbal measure has the best psychometric characteristics and the analytical measure the least good. The analytical measure has the greatest percentage of large gains and large decreases, and these gains are not fully explained by the lower reliability of the measure. The advantage provided by taking the same edition twice rather than two different editions is greatest for the analytical measure. The test-retest correlations for the analytical measure are considerably lower than those for the other measures. On the positive side, the average gain for repeaters taking different editions of the test is somewhat lower than that for the other two measures.

The verbal measure, on the other hand, has the smallest percentage of large score changes. Taking the same edition twice provides less advantage on the verbal measure than on the other two measures. Also, the test-retest correlations are the highest of the three measures. The psychometric characteristics with regard to score changes for the quantitative measure are between those for the verbal and analytical measures.

The lower average gain for the analytical measure is consistent with the elimination of item types susceptible to short-term practice effect. Why the analytical measure has the other properties that were observed is a matter for conjecture. For most of the analytical reasoning and logical reasoning items, several items are based on a single stimulus (explanatory passage). For analytical reasoning, in particular, encoding and conceptualizing the information presented in the stimulus material is a major aspect of determining the correct response to the items that follow the stimulus. Stewart (1983) has hypothesized that this might result in a factor structure in which there was a large specific variance due to particular stimulus material. This might explain several of the reported findings.

When examinees repeat the same edition of the test, the stimulus-specific variance for the analytical measure is shared across the two administrations because the factorial structure is identical. Thus, the gain for the analytical measure is greater than that for the verbal and quantitative measures. When two different editions of the test are administered, only the variance from the general analytical factor(s)[1] is shared. Therefore, in terms of the stimulus-specific variance, it is as if the second administration had not been preceded by the first administration of the measure. Thus, there is less gain, on the average, for the analytical measure. This same reasoning, however, would lead one to expect a high test-retest correlation for repeaters taking the same edition of the analytical measure twice and a lower correlation for those taking different editions. Although the latter expectation was confirmed, the former was not. That is, although the same-edition correlation was considerably higher than the different-edition correlation, it was still considerably lower than the same edition correlation for the verbal and quantitative measures. Perhaps the effect of the shared stimulus-specific variance was ameliorated by the relatively high scaled-score standard error of measurement due to the shortness (50 items) of the analytical measure.

The correlations for examinees taking two different editions of the quantitative measure, although higher than those for the analytical measure, are lower than the correlations for the verbal measure. A possible explanation for this is that the GRE quantitative measure is more amenable to improvement from short term (several months or less) studying than either the verbal or the analytical measure (Powers, 1983). Thus, individual variation in preparing to repeat the General Test might be magnified for the quantitative measure and therefore lead to a lower correlation. Livingston (1984) suggests that the study of mathematics is more likely to generalize across problems than study of vocabulary is to generalize across words and relationships. In mathematics, one new concept allows you to solve several different problems. In vocabulary, one new concept gets you one new word - maybe two. Hence, the effects of study might be greater for quantitative than for verbal scores. This is less likely to be a factor for examinees taking the same edition of the test twice, since it is relatively easy to learn the answers to specific, known questions for each of the three measures.

---

[1]Much evidence exists that these general analytical factors might be a verbal and a quantitative factor (Kingston, 1984).

REFERENCES

Goodison, M., & Wild, C. (1982). Evaluation of the Graduate Record Examinations (GRE) General (Aptitude) Test 1981-82. Unpublished paper. Princeton, NJ: Educational Testing Service.

Kingston, N. (1984). Reanalysis of the psychometric characteristics of the revised analytical measure of the GRE General Test. Unpublished paper. Princeton, NJ: Educational Testing Service.

Kingston, N., & Dorans, N. (1982) The effect of the position of an item within a test on item responding behavior: an analysis based on item response theory. GRE Board Professional Report 79-12Pb. Princeton, NJ: Educational Testing Service.

Livingston, S. (February 1984). Personal communication.

Lord, F., & Novick, M. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing Company.

Powers, D. (1983). Effects of coaching on GRE Aptitude Test scores. GRE Board Research Report 81-3R. Princeton, NJ: Educational Testing Service.

Rock, D., & Werts, C. (1980). An analysis of time-related score increments and/or decrements for the GRE repeaters across ability and sex groups. GRE Board Research Report 77-9R. Princeton, NJ: Educational Testing Service.

Stewart, E. E. (October 1983). Personal communication.

Swinton, S., Wild, C., & Wallmark, M. (1983). Investigation of practice effects onitem types in the Graduate Record Examinations Aptitude Test. GRE Board Professional Report 80-1cP. Princeton, NJ: Educational Testing Service.

Wallmark, M. (1982a). GRE test analysis - Aptitude Test form 3DGR1. Unpublished Statistical Report 82-35. Princeton, NJ: Educational Testing Service.

Wallmark, M. (1982b). GRE test analysis - Aptitude Test form 3DGR3. Unpublished Statistical Report 82-64. Princeton, NJ: Educational Testing Service.

Wallmark, M. (1982c). GRE test analysis - Aptitude Test form 3DGR2. Unpublished Statistical Report 82-76. Princeton, NJ: Educational Testing Service.

Wallmark. M. (1982d). GRE test analysis - Aptitude Test form 3EGR1. Unpublished Statistical Report 82-81. Princeton, NJ: Educational Testing Service.

Wallmark, M. (1982e). GRE test analysis - Aptitude Test form EGR2. Unpublished Statistical Report 82-83. Princeton, NJ: Educational Testing Service.

APPENDIX 1

DEMOGRAPHIC CHARACTERISTICS OF GRE REPEATERS

In analyzing the data from 11,509 candidates who had taken the GRE at least twice during the two year period, an effort was made to determine whether the repeaters differed substantially from a recent group of nonrepeaters in terms of demographic characteristics. A group of 200,361 examinees who took the GRE between October 1981 and April 1982 was identified as a control group. Eighty-seven percent of the examinees in the 1981-82 GRE population indicated that they had never taken the GRE before, and an additional 11 percent said they had taken one or more GRE tests prior to October 1981. Therefor, the GRE 1981-82 population may be regarded as primarily a group of nonrepeaters, and its overlap with the repeater group in question is certainly minimal.

Because of the disparity in the sizes and response rates of the repeater group and the 1981-82 GRE population (henceforth termed the nonrepeater group), adjusted percentages will be used for the comparison of responses to the background questions. The percentages presented exclude nonresponses, which were more prevalent in the nonrepeater than in the repeater group (on average, about 23%, as compared to about 12% for the repeaters). The lower nonresponse for the repeaters is not necessarily due to their greater diligence in answering background questions but rather to the fact that the questionnaire responses are updated on the candidates' records after each administration.

The following discussion of demographic variables will refer to the adjusted percentages in Table A1-1.

Sex. The adjusted percentages of males and females in the repeater population differed very little from those in the primarily nonrepeating 1981-82 GRE examinee population.

Citizenship. The percentage of U.S. citizens among repeaters (79.9%) was somewhat lower than among the nonrepeaters (86.3%). The repeater population consisted of a higher percentage of resident aliens (13.1%) than did the nonrepeater population (3.6%). However, there was a lower percentage of repeaters(6.9%) than nonrepeaters (10.1%) who considered themselves neither citizens nor aliens. This is due to the fact that 600 repeaters who had taken foreign administrations were dropped from the current data set whereas the GRE 1981-82 examinee data set included candidates taking overseas administrations. Had the foreign administration repeaters been included, they would have diminished the percentage of U.S. citizens to about 75 percent while augmenting the "neither" category by the same amount (given the exclusion of the foreign administration records, it is difficult to determine the status of those repeaters who are neither citizens nor resident aliens, although they might be assumed to be non-resident aliens).

Ethnic group. Blacks, Orientals, Hispanics and other non-Whites all had greater representation in the repeater (25.1%) than in the nonrepeater group (18.1%). It is possible that these minority or foreign examinees may choose to

repeat the GRE in the hope of having somewhat greater verbal facility or familiarity with standardized tests by the time of the second administration.

English as a primary language. A higher percentage of repeaters (16.9%) than nonrepeaters (10.7%) indicated that English was not their primary language. English-secondary examinees, especially those who are newly arrived in the United States, may hope that their second administration scores will reflect their growing facility in the language.

Graduate degree objective. The repeater group contained a higher percentage of examinees with Ph.D. aspirations (42.6%) than did the nonrepeater group (35.4%). Conversely, the percentage of repeaters with master's degree objectives (52.1%) was lower than that in the nonrepeater group (58.8%). Since doctoral programs apply more stringent admission requirements for their applicants, it is not surprising that doctoral aspirants would have a greater desire to enhance their scores than would examinees seeking a master's or intermediate degree.

Self-reported undergraduate major GPA and self-reported undergraduate GPA in the last two years. The repeaters tended to have slightly lower GPAs than did the nonrepeaters. Whereas 40.4 percent of the repeaters claimed undergraduate major GPAs of A or A-, 45.3 percent of the nonrepeaters claimed such high GPAs. In the case of undergraduate GPA in the last two years, 43.2 percent of the repeaters claimed averages of A or A- while 48.0 percent of the nonrepeaters claimed such averages. The repeaters, who showed higher concentrations of B, B-, and C grades, may have felt compelled to retake the GRE in hopes of higher scores that might compensate for their relatively low GPAs. In addition, the lower repeater undergraduate GPAs, along with the lower repeater GRE scores, may indicate that the repeaters are a less able subset of the GRE examinee population.

Educational level at the time of the first administration. The repeaters tended to be less advanced in their educational level than the nonrepeaters. Whereas 58.2 percent of the repeaters were undergraduates (sophomores, juniors, or seniors) when they took their first GRE test, only 43.8 percent of the GRE 1981-82 examinee population fell into these categories. The repeater group contained lower percentages of more advanced students (unenrolled bachelor's recipients, unenrolled master's recipients, first and second year graduate students, and "others") than did the nonrepeater group. Since the majority of the repeaters were undergraduates at the time of their first administration, they may have felt less pressed to submit their first-time scores than would more advanced students who are taking the test to satisfy immediate graduate admissions or other requirements.

Type of intended graduate major. For both the repeater and nonrepeater groups, codes of intended graduate major were grouped into four broad catagories: humanities, social sciences, physical sciences, and biological sciences. The percentages of repeaters in the social and physical sciences are roughly comparable to those of the nonrepeaters. However, the percentage of repeaters intending humanities majors (9.1%) is slightly lower than the 13.2 percent of nonrepeaters intending to enroll in humanities departments. Conversely, 25.0 percent of the repeaters intended to apply to biological sciences departments, as opposed to 20.6 percent of the nonrepeaters. In some cases, the repeaters

may be applying to departments whose admission requirements are quite stringent, and so these examinees feel compelled to achieve higher scores than they obtained during their first administration.

Specific intended graduate majors. Only 6.0 percent of the repeaters intended to apply to education departments while 8.6 percent of the nonrepeaters intended to major in education. On the other hand, higher percentages of repeaters than nonrepeaters intended to apply to psychology, veterinary medicine, electrical engineering, microbiology and geology departments. The fact that 14.0 percent of the repeaters intended to apply to psychology departments (as opposed to 9.0 percent of nonrepeaters) seems notable. These aspiring psychologists are probably convinced not only of their ability to do better the second time around but also of their need for high scores to improve their chances of admission to a selective institution. Indeed, the selectivity of institutions offering the intended graduate major is probably an incentive for students to repeat the GRE. Conversely, fields like education, which in most institutions tend to admit the majority of their applicants, probably do not create incentives to repeat.

Summary. In comparing the repeater group to the group of 200,361 primarily nonrepeating examinees, the following factors were associated with retaking the GRE:

-- Noncitizenship
-- English as a secondary language
-- Non-White ethnic status
-- Ph.D. aspirations
-- Lower-than-usual self-reported undergraduate GPAs
-- Less advanced educational level (primarily undergraduate)
-- Intentions to apply to more selective department types

Table A1-1
Frequencies and Adjusted Percentages* of Repeaters and Nonrepeaters on Demographic Variables

Variable	Repeaters		Nonrepeaters	
	N	Adjusted %	N	Adjusted %
SEX				
Male	5497	47.8%	95707	48.3%
Female	6006	52.2%	102546	51.7%
CITIZENSHIP				
U.S. citizen	8217	79.9%	139752	86.3%
Resident alien	1351	13.1%	5856	3.6%
Neither	710	6.9%	16302	10.1%
ETHNIC GROUP				
White	7086	74.8%	120928	81.8%
Black	715	7.5%	9619	6.5%
Oriental	547	5.8%	5341	3.6%
Hispanic	487	5.1%	5691	3.8%
Other	636	6.7%	6292	4.2%
ENGLISH PRIMARY				
Yes	8487	83.1%	143178	89.3%
No	1730	16.9%	17079	10.7%
GRADUATE DEGREE OBJECTIVE				
Nondegree	32	0.3%	1325	0.8%
Master's	5297	52.1%	93464	58.8%
Intermediate	166	1.6%	3074	1.9%
Ph.D.	4331	42.6%	56289	35.4%
Postdoctorate	347	3.4%	4818	3.0%
SELF-REPORTED UNDERGRADUATE MAJOR GPA				
D or lower	13	0.1%	129	0.1%
C-	33	0.3%	532	0.3%
C	632	6.3%	8501	5.5%
B-	1582	15.8%	21104	13.6%
B	3695	37.0%	54916	35.3%
A-	2727	27.3%	45729	29.4%
A	1309	13.1%	24770	15.9%

*Percentages are adjusted for nonresponse, which varies by question, but averages about 23 percent for nonrepeaters and 12 percent for repeaters.

Table A1-1 Continued

Variable	Repeaters N	Repeaters Adjusted %	Non repeaters N	Non repeaters Adjusted %
SELF-REPORTED LAST 2 YEARS GPA				
D or lower	17	0.2%	139	0.1%
C-	30	0.3%	478	0.3%
C	515	5.2%	7139	4.6%
B-	1392	14.0%	18666	12.0%
B	3707	37.2%	54326	34.9%
A-	3070	30.8%	51830	33.3%
A	1241	12.4%	22889	14.7%
EDUCATIONAL LEVEL AT FIRST ADMINISTRATION				
Sophomore	52	0.5%	244	0.1%
Junior	728	6.5%	3008	1.5%
Senior	5736	51.2%	82109	42.2%
Graduate student Year 1	1032	9.2%	23047	11.8%
Graduate student Year 2	507	4.5%	8395	4.3%
B.A., unenrolled	1704	15.2%	47948	24.7%
M.A., unenrolled	795	7.1%	16900	8.7%
Other	645	5.8	12837	6.6%
TYPE OF INTENDED GRADUATE MAJOR				
Humanities	839	9.1%	18325	13.2%
Social sciences	4276	46.4%	64712	46.5%
Biological sciences	2306	25.0%	28701	20.6%
Physical sciences	1793	19.5%	27554	19.7%
SPECIFIC INTENDED GRADUATE MAJORS				
Education	555	6.0%	11981	8.6%
Psychology	1292	14.0%	12503	9.0%
Biology	140	1.5%	2039	1.5%
Microbiology	163	1.8%	1305	0.9%
Veterinary medicine	306	3.3%	2061	1.5%
Chemistry	143	1.5%	2481	1.8%
Electrical engineering	249	2.7%	3538	2.5%
Geology	252	2.7%	2736	2.0%
Computer science	300	3.3%	5411	3.9%

A special repeater subgroup: Examinees with scores of 200 on the verbal, quantitative, or analytical measure for either administration. A subgroup of repeaters with scores of 200 on the verbal, quantitative, or analytical measure for either the first or the second administration was extracted for special consideration. Five-hundred-and-sixty-one examinees with at least one 200 score were identified, although some of these may have had more than one 200 score. The records may be broken down as follows:

Table A1-2

Number of Repeaters Scoring 200
on Each Measure of the GRE General Test

Administration	Test Name	Number of 200 Scores
1	Verbal	228
1	Quantitative	87
1	Analytical	88
2	Verbal	152
2	Quantitative	78
2	Analytical	40

It was not surprising to find that, for each test, the number of 200 scores for the first administration was larger than that for the second administration. Also unsurprising, the verbal test elicited more scores of 200 than did the quantitative or analytical tests, especially when the demographic characteristics of the 200 scorers were considered.

Demographic characteristics of total group of 200 scorers. This group had the following characteristics (with percentages adjusted for missing data): (Percentages for the second administration test of each type are not reported since they closely resemble those for the first administration.)

Verbal test 200 scorers

- 80 percent were male;
- 87 percent were noncitizens;
- 74 percent were non-White;
- 70 percent did not consider English their primary language; and
- 81 percent had master's degree aspirations.

Quantitative test 200 scorers

- 69 percent were female;
- 79 percent were citizens;
- 68 percent were non-White (46% were Black);
- 78 percent considered English their primary language; and
- 67 percent had master's degree aspirations.

Analytical test 200 scorers

- 76 percent were male;
- 77 percent were noncitizens;
- 67 percent were non-White;
- 57 percent did not consider English their primary language; and
- 75 percent had master's degree aspirations.

In summary, the profile of a person who achieves a score of 200 on the verbal test fairly closely resembles that of a 200-scoring analytical test examinee (i.e., he is a male, non-White, English-secondary language, noncitizen with master's degree aspirations). However, the typical examinee scoring 200 on the quantitative test is quite different (i.e., she is a female, non-White, English-primary language, citizen with master's degree aspirations). It was not possible to determine from the examinee records themselves whether the examinees achieved scores of 200 through lack of knowledge of the subject matter tested, through random marking of answer sheets, or through simple omission of the entire measure.

APPENDIX 2

DIFFERENCES WITH REGARD TO BACKGROUND VARIABLES OF THREE GROUPS OF REPEATERS

To determine whether or not the three repeater groups differed in their composition, comparisons of groups (New-New, Old-Old, and Old-New) by background variables were obtained. Results of these comparisons are presented in Table A2-1. The percents in parentheses represent the number of observations in the cell divided by the number of respondents for the particular variable. The percents were adjusted for missing values because the New-New group had a higher percentage of missing cases than the other two groups (missing cases averaged about 17% across variables in the New-New group as compared to about 10% in the Old-Old and Old-New groups). Missing data were due primarily to nonresponse, except in the case of the age variable, for which values of 18 or less and values of 64 or more were counted as missing (the numbers of cases in these categories were not considered believable), and in the case of the educational level variable, for which sophomores and "others" were combined with the nonresponses by reason of low cell counts.

Age at the time of the first administration was obtained by subtracting the examinee's birth year from the year in which the first administration occurred. The Old-New group showed a somewhat lower percentage of 19- to 21-year olds and a slightly higher percentage of 22- to 24- year olds than did the other two groups. The New-New and Old-Old groups seemed roughly comparable in terms of the percentage of observations in each age category, although there was a slightly lower percentage of students over 30 years old in the Old-Old group.

Educational level at the time of the first administration was obtained from the registration form. Once again, the New-New and Old-Old groups were roughly comparable in each category. The Old-New group, however, contained a higher percentage of juniors and a lower percentage of seniors than did the other two groups. The Old-New group also contained a higher percentage of unenrolled bachelor's degree recipients but a lower percentage of unenrolled master's degree recipients than did the other two groups.

Type of intended graduate major was obtained by grouping specific examinee-designated department codes into four broad categories (humanities, social sciences, biological sciences, and physical sciences). Examinees who did not specify a department code or who specified a code of "other" were counted as missing. As with the other variables, the cell percentages were adjusted for missing cases since the proportion of missing cases in the New-New group was substantially higher than those in the other two groups. The percentages of New-New group candidates who intended graduate majors in the social sciences and biological sciences were somewhat lower, and the percentages who planned to enter humanities and physical sciences departments were somewhat higher, than in the other two groups.

Sex, which is entered on the registration form, was specified by almost every examinee. The relative percentages of males and females were approximately the same across all three groups.

Citizenship was obtained from the background questionnaire and shows a few differences among the three groups. The percentage of U.S. citizens was slightly lower, and the percentages of candidates who considered themselves neither citizens nor resident aliens somewhat higher, in the New-New group than in the other two groups. There appears to be a slight tendency toward more noncitizens taking the more recent version of the GRE, which may reflect the growing foreign student enrollments in American graduate schools.

Ethnic status shows a trend similar to that of the citizenship variables. As the percentage of White candidates decreases slightly in going from Old-Old administrations through Old-New administrations to New-New administrations, the percentages of Blacks, Oriental, Hispanics, and others all tend to rise slightly.

English as a primary language also reflects the influx of foreign and minority candidates into the repeater population. In the New-New group, 19.4 percent of the examinees claimed to communicate best in some language other than English as opposed to 16.0 percent in the Old-Old group and 15.0 percent in the Old-New group.

Graduate degree objective also revealed some slight differences among groups, particularly in terms of relative proportions of candidates with master's and doctoral aspirations. Only 40.8 percent of the New-New group candidates specified a Ph.D. objective, as opposed to 42.1 percent of the Old-Old group and 46.3 percent of the Old-New group. On the other hand, 54.5 percent of the New-New group indicated a master's objective, as opposed to 52.3 percent in the Old-Old group and 47.9 percent in the Old-New group. While these differences are not striking, they may reflect a slight tendency toward lowered degree aspirations among the more recent GRE examinees.

Last two years undergraduate GPA results may reflect some of the tendencies we have seen in other variables. The New-New group had slightly higher percentages of candidates with B or B- averages and slightly lower percentages with A or A- averages than the Old-Old and Old-New groups. The somewhat lower GPAs may have some relationship to the lower degree aspirations, lesser English-speaking ability, and greater minority composition of the New-New group.

Summary. Although there were no truly substantial differences among the three groups for any of the background variables examined, there were some indications that the New-New group differs slightly from the Old-Old and Old-New groups. These may be summarized as follows:

-- The New-New group had a higher percentage of missing cases on most of the variables examined. Since the remaining samples in each group were adjusted for these missing cases, the results may be slightly contaminated by the higher level of nonresponse in the New-New group.

-- The New-New group had a lower percentage of intended social sciences majors and a higher percentage of intended humanities and physical sciences majors than the other two groups.

-- The New-New group had a higher percentage of noncitizens, non-Whites and non-native English speakers than the other two groups.

-- The New-New group had a higher percentage of examinees with master's aspirations and a lower percentage with doctoral aspirations than the other two groups.

-- The New-New group had a higher percentage of examinees with B or B- last two-years undergraduate GPAs and a lower percentage with A or A- GPAs than the other two groups.

Table A2-1
Comparisons of Repeater Groups (New-New, Old-Old, Old-New)
by Background Variables*

	New-New	Old-Old	Old-New
AGE AT FIRST ADMINISTRATION			
19-21	1203 (29.7%)	1566 (31.1%)	529 (22.7%)
22-24	1237 (30.5%)	1561 (31.1%)	858 (36.8%)
25-29	881 (21.7%)	1053 (20.9%)	514 (22.1%)
30-35	449 (11.1%)	530 (10.6%)	274 (11.8%)
36-63	280 (6.9%)	316 (6.3%)	153 (6.6%)
EDUCATIONAL LEVEL AT FIRST ADMINISTRATION			
Junior	100 (2.7%)	125 (2.7%)	503 (23.7%)
Senior	2161 (57.8%)	2768 (59.7%)	807 (38.0%)
B.A. (unenrolled)	569 (15.2%)	750 (16.2%)	385 (18.1%)
Graduate Student year 1 or 2	583 (15.6%)	683 (14.7%)	273 (12.8%)
M.A. (unenrolled)	325 (8.7%)	312 (6.7%)	158 (7.4%)
TYPE OF INTENDED GRADUATE MAJOR			
Humanities	311 (10.1%)	383 (9.2%)	145 (7.3%)
Social Sciences	1371 (44.6%)	1965 (47.2%)	940 (47.5%)
Biological Sciences	746 (24.3%)	1024 (24.6%)	536 (27.1%)
Physical Sciences	644 (21.0%)	792 (19.0%)	357 (18.0%)
SEX			
Male	1931 (47.3%)	2453 (48.4%)	1113 (47.3%)
Female	2151 (52.7%)	2615 (51.6%)	1240 (52.7%)

*Percents represent N in cell / [(N in relevant group) - (N missing on variable)]
(N for New-New = 4084, N for Old-Old = 5072; N for Old-New = 2353).

Table A2-1 Continued

	New-New	Old-Old	Old-New
CITIZENSHIP			
U.S. Citizen	2665 (77.3%)	3749 (80.7%)	1803 (82.6%)
Resident Alien	304 (8.8%)	829 (17.8%)	218 (10.0%)
Neither	480 (13.9%)	69 (1.5%)	161 (7.4%)
ETHNIC STATUS			
American Indian	30 (0.9%)	36 (0.9%)	10 (0.5%)
Black	263 (8.2%)	281 (6.6%)	171 (8.4%)
Oriental	223 (7.0%)	222 (5.2%)	102 (5.0%)
Hispanic	175 (5.5%)	204 (4.8%)	108 (5.3%)
White	2312 (72.3%)	3239 (76.5%)	1535 (75.3%)
Other	196 (6.1%)	252 (6.0%)	112 (5.5%)
COMMUNICATE BEST IN ENGLISH			
Yes	2748 (80.6%)	3888 (84.0%)	1851 (85.0%)
No	661 (19.4%)	743 (16.0%)	326 (15.0%)
GRADUATE DEGREE OBJECTIVE			
Nondegree	8 (0.2%)	20 (0.4%)	4 (0.2%)
Masters	1851 (54.5%)	2409 (52.3%)	1037 (47.9%)
Intermediate	45 (1.3%)	79 (1.7%)	42 (1.9%)
Ph.D.	1388 (40.8%)	1941 (42.1%)	1002 (46.3%)
Postdoctorate	107 (3.1%)	159 (3.5%)	81 (3.7%)
LAST 2 YEARS UNDERGRADUATE GPA			
C or lower	191 (5.7%)	257 (5.7%)	114 (5.3%)
B-	478 (14.4%)	608 (13.5%)	306 (14.3%)
B	1278 (38.4%)	1664 (36.9%)	765 (35.8%)
A-	977 (29.3%)	1415 (31.4%)	678 (31.7%)
A	405 (12.2%)	560 (12.4%)	276 (12.9%)

APPENDIX 3

REGRESSION EQUATIONS

Table A3-1

Regression Equation for the Second Verbal Score[1]

Variable	B Weight	Beta Weight
First verbal score (V)	9.3×10^{-1}	.87
First V^{-3}	4.9×10^{8}	.10
First V^{3}	-1.4×10^{-7}	-.08
First analytical score	2.0×10^{-1}	.21
Does not communicate better in English	-1.3×10^{1}	-.04
Birth year: 1931 to 1935	2.5×10^{1}	.02
Birth year: 1941 to 1945	1.5×10^{1}	.02
Years since B.A.: 25 to 45	3.5×10^{1}	.02
Degree objective: Ph.D.	7.0×10^{0}	.03
Neither citizen nor resident alien	-1.9×10^{1}	-.02
Major: biological science	5.4×10^{0}	.02
Ethnic group: Other	-9.7×10^{0}	-.02
Undergraduate school: Private, not church affiliated	5.4×10^{0}	.02
Constant	-4.1×10^{1}	

Multiple r = .9037

[1] Based on stepwise multiple regression in a group of 4,075 examinees who took the pre-October 1981 form of the GRE General Test both times.

Table A3-2

Regression Equation for the Second Quantitative Score[1]

Variable	B Weight	Beta Weight
First quantitative score	7.1×10^{-1}	.69
First analytical score	1.2×10^{-1}	.12
First verbal score	5.4×10^{-2}	.05
Major: physical science	1.8×10^{1}	.05
Major: social science	-1.7×10^{1}	-.05
Major: humanities	-1.7×10^{1}	-.04
Major: psychology	-1.5×10^{1}	-.04
Major: education	-1.6×10^{1}	-.03
Major: physical education	-2.2×10^{1}	-.02
United States citizen	-2.0×10^{1}	-.07
Female	-1.5×10^{1}	-.06
Ethnic group: Black	-1.5×10^{1}	-.03
Ethnic group: Oriental	1.3×10^{1}	.02
Ethnic group: Hispanic other than Puerto Rican or Mexican-American	-1.9×10^{1}	-.02
Birth year: 1941 to 1945	-1.8×10^{1}	-.03
Birth year: 1960	1.1×10^{1}	.02
Years since B.A.: one or less	5.0×10^{0}	.02
Years since B.A.: 15 to 19	2.1×10^{1}	.02
Does not communicate better in English	6.6×10^{0}	.02
Undergraduate school: private, not church affiliated	5.5×10^{0}	.02
Undergraduate grade-point average: A	9.1×10^{0}	.02
Undergraduate grade-point average: A-	5.5×10^{0}	.02
Constant	1.2×10^{2}	

Multiple r = .8657

[1] Based on stepwise multiple regression in a group of 4,075 examinees who took the pre-October 1981 form of the GRE General Test both times.